MO11 Avoiding Problems in CMOS Memory Operation

INTRODUCTION

This application note recognizes two problems often encountered when CMOS memories are designed and operated in low power digital systems. The causes and effects of these problems are analyzed and a few solutions or methods of avoiding them are put forth. The emphasis in this note will be on a throughly detailed presentation of the problems, since CMOS memories are now used in quite a wide variety of applications and the solutions documented here may not be feasible or appropriate for all of them. With a clear understanding of the problems it is hoped that system engineers will be able to formulate their own solu- ons, particularly suited to their unique resources and constraints.

INPUT OVER-VOLTAGE AND SCR LATCHUP

The Problem

SCR latchup is a phenomenon inherent in the basic CMOS process structure. It is brought about by external conditions, perhaps present only momentarily, and once induced is difficult to reverse except by complete removal of power to the chip. Latchup results in a large current flow from V_{CC} to GND, typically exceeding 100mA, and limited only by the external V_{CC} power supply limit. If there is no limit, or if the limit is too high, the current can cause internal bond wires to melt and open thereby disconnecting the silicon die from the package leads. Obviously this can cause severe system problems such as blown fuses, excessive heat, power supplies pulled low, and prematurely drained batteries, as well as failures of individual devices, which can be electrically modified or destroyed. In the following discussion we will look closely at SCR latchup in RAMs and other CMOS memories.

Over-voltage at Inputs vs. Over-voltage at V_{CC}

Much has been written about PNPN latchup since it was first observed as a technical problem in commercially available CMOS circuits. Engineers at RCA characterized it as it applied to the popular 4000 series SSI and MSI circuits[1], and since that time, occasionally papers have appeared that contribute to the general understanding of this problem. However, analysis thus far has centered on the general aspect of latchup, that is, the two-terminal PNPN, or Shockley diode (fig. 1(a)). In circuit applications, the two-terminal analysis represents latchup as being triggered by an increasing voltage across the power terminals, such as an excessive voltage at the V_{CC} pin (normally well above the maximum V_{CC} rating of the device). While this is a valid latchup mechanism, which can occur in a system if large transients are experienced on the V_{CC} line during normal operation or power-up, it is easily prevented by adequate system power regulation.

Unfortunately, even though system power supplies are well-regulated, inputs to individual CMOS circuits often are not. Latchup is also known to occur in CMOS systems when voltage applied to an input exceeds the voltage at V_{CC} by some fixed amount. Input over-voltage can be caused by transients on the input lines of a single circuit board, lines of a single circuit board, overdriving during V_{CC} power-up or power-down cycles, or by inputs at high voltage levels on other system circuit boards. For this reason we will look at the three terminal PNPN device (fig. 1(b)). This represents the PNPN structure between V_{CC} and GND with an input or "gate" so placed as to enhance the likelihood of a latchup event. The gate in our analysis is provided by one or more inputs to the CMOS circuit. *The presence of the input "gate" may enable latchup even though V_{CC} over-voltage conditions are not present.* The three terminal device is known as an SCR or thyristor; it reduces to the two terminal characteristic if the gate current becomes zero.

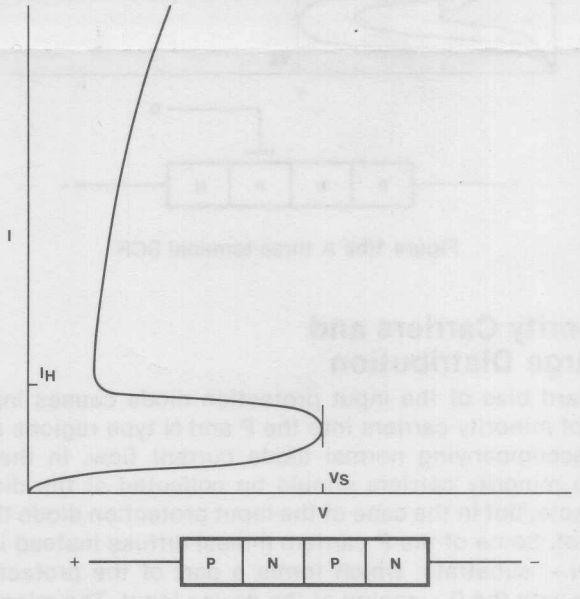


Figure 1(a): A two-terminal PNPN diode

A Three-terminal PNPN

To properly examine the effects of over-voltage at CMOS device inputs, we must consider the basic bulk CMOS structure, see (fig. 2). The PNPN structure, shaded in 2 (a), is made up of the highly doped source and drain regions of a complementary transistor pair and the lightly doped P – well and N – substrate. The transistor pair is typical of inverters anywhere on the chip. The shaded path is essentially the two-terminal device of 1 (a).

[1] Flood, J. and Pujol, H.L., "Guide to Better Handling and Operation of CMOS Integrated Circuits", RCA COS/MOS Integrated Circuits, 1977

INTERSIL, INC., 10710 N. TANTAU AVE., CUPERTINO, CA 95014 (408) 996-5000 TWX: 910-338-0171

The third terminal is provided by a part of the CMOS structure not even related to normal operation of the transistor pair. It is a PN junction apart from the inverter, commonly called an input protection diode. This diode serves to clamp high voltages caused by ordinary electrostatic discharges and prevents them from damaging the delicate polysilicon gates of the transistors. The input protection diode is shown in relation to the transistors in fig. 2 (b).

Useful for protecting the internal CMOS circuits, particularly when power is not applied to the chip, the input protection diode can become forward biased under normal operating conditions if an over-voltage of sufficient magnitude is applied to a protected input. When this happens an effective SCR gate is created, and conditions present can enable SCR action.

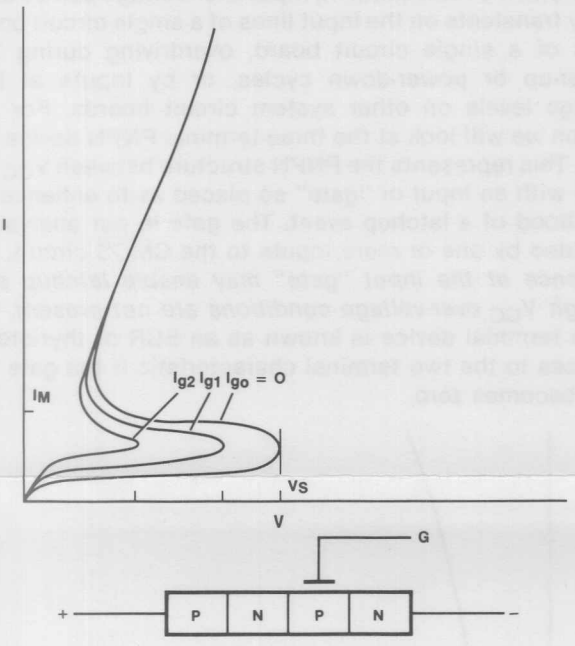


Figure 1(b): A three-terminal SCR

Minority Carriers and Charge Distribution

Forward bias of the input protection diode causes injection of minority carriers into the P and N type regions and the accompanying normal diode current flow. In theory these minority carriers should be collected at the diode contacts, but in the case of the input protection diode they are not. Some of the P carriers (holes) diffuse instead into the N– substrate, which forms a part of the protection diode with the P+ region of the device input. The minority carrier holes make up a charge distribution in the substrate.

The holes that make up this distribution recombine with electrons in the substrate when the input protection diode is marginally forward-biased. As the voltage at the input is increased, and as current through the diode increases, the charge distribution in the substrate also increases. When the hole charge distribution reaches a certain level, some of the holes do not recombine with electrons, but diffuse to the boundary of the reverse-biased P–N– junction formed by the P– well of an N–channel transistor and the substrate. The reverse-biased junction acts as a collector for the minority holes, and the holes drift across it into the P– well. In the P– well the holes are majority carriers not subject to recombination with electrons. They are "free" to flow as small currents in the P– well. The activity described above is roughly diagrammed in fig. 3.

The PN junction of the protection diode acts as a capacitor when the voltage at the input increases; that is, voltage does not appear across it instantaneously. Increased current flows through the junction as in a capacitor and the minority charge distribution on either side is increased. The build-up of charge distribution is time dependent on changing voltage, and for this reason there is a relationship between the magnitude of the input over-voltage and the time to latchup. If the over-voltage is very large, a large current flows through the substrate junction and the minority holes in the substrate increase rapidly. If the over-voltage is smaller, charge build-up in the substrate takes more time. We will look at the relationship in terms of time delay to latchup in a later section.

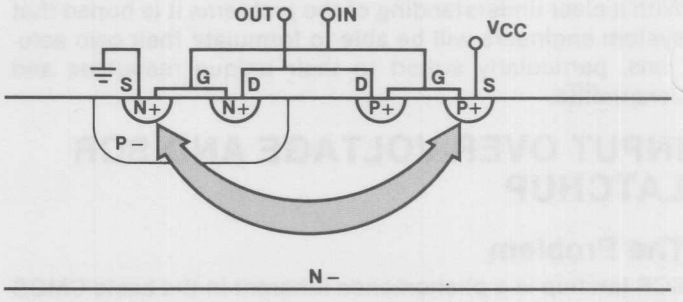
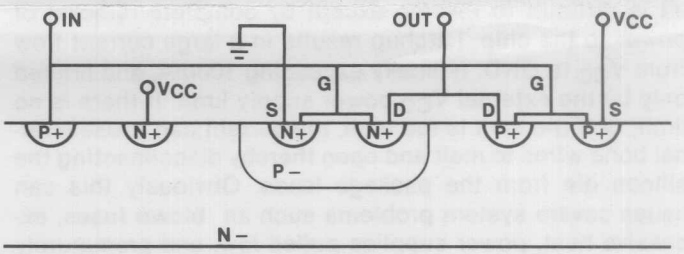


Figure 2 (a): PNPN structure in bulk CMOS



(b): Shown with input protection diode

The Latchup Phenomenon

The excess holes collected into the P– well form the gate or trigger current of the three-terminal PNPN device. Their effect can be demonstrated with the commonly accepted two transistor model of fig. 4.

Current in the P– well, made up of holes, causes a voltage drop in the direction of its flow. The voltage drop is the product of the current and the resistance of the current path, determined by the resistivity of the P– well material. At some point the voltage drop causes a large enough difference in potential between the outer region of the P– well and the N+ source region to forward bias the P–N+ junction. Analogous to a diode drop across R_p in fig. 4, the difference in potential turns on a latent NPN transistor (Q_1) and enables amplified electron current to flow into the substrate.

The electrons collected in the substrate are majority carriers and do not recombine with holes. They, too, form a current in uniformly resistive material, and cause a drop in potential in the direction of a P+ contact (normally a source for a P–channel transistor). The voltage drop

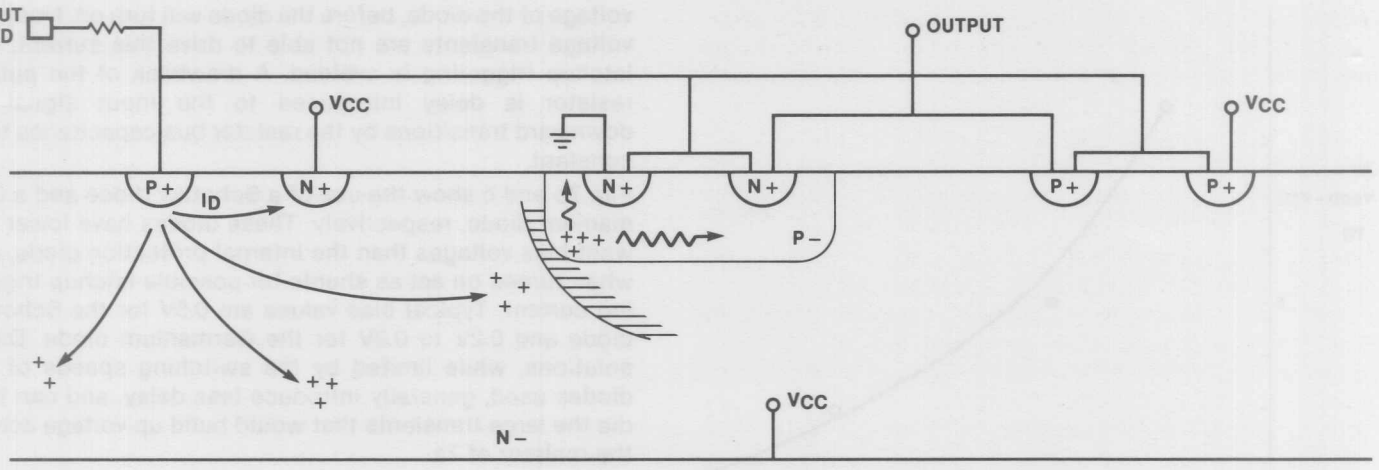


Figure 3: Minority carrier activity due to input over-voltage

allows forward biasing of the N – P + junction, and the latent PNP transistor (Q_2) is activated to supply more holes to the P – well as collector current. Effectively, the required voltage drop is established across R_n of fig. 4. Current flow is regenerative when the holes supplied by collector current from the PNP contribute to the hole current originally collected from the input protection diode; or, more simply, when the product of transistor betas is one or greater. In a very short time, a large current is flowing through the PNPN structure from V_{CC} to GND.

Time Dependent Effects

As is mentioned above, the delay to latchup due to over-voltage at a protected input is dependent on the minority carrier charge distribution, and from that the amplitude of the over-voltage. The amplitude of the over-voltage determines current through the protection diode, and the rate of increase of hole charge distribution in the substrate.

The amplitude to delay relationship varies from input to input on a single chip and from device to device in a single lot. The precise relationship is determined by a number of factors- proximity of the input protection diode to circuit elements, resistivity of the P – and N – material, diode threshold levels, value of the input resistor and others. Mathematical calculations are made extremely difficult by these variables, therefore it must suffice to explore these relationships on the basis of empirical data.

Fig. 5 shows a curve extrapolated from data taken on a IM65x08 (1kx1) CMOS RAM. The data confirms the expected decline in over-voltage necessary to activate latchup with increasing time of application. The curve in fig. 5 approached a DC limit of 0.9v; below this level latchup did not occur on the tested device, even after several seconds. Over voltage is plotted as DV, the difference between voltages at address inputs and V_{CC}.

The French Centre National D'Etudes Spatiales conducted an experiment of this nature with several CMOS RAMs, obtaining the results plotted in fig. 6[2]. The devices tested were a 65x08 similar to the device mentioned above, an Intersil 6508 fabricated with a bulk silicon process, and other 1k and 4k CMOS RAMs. The Intersil bulk silicon process, which since 1979 has been replaced by the smaller-geometry, oxide-isolated SELOX-C® process, clearly offered the best resistance to latchup. The difference is marked, and systems originally configured with the IM65

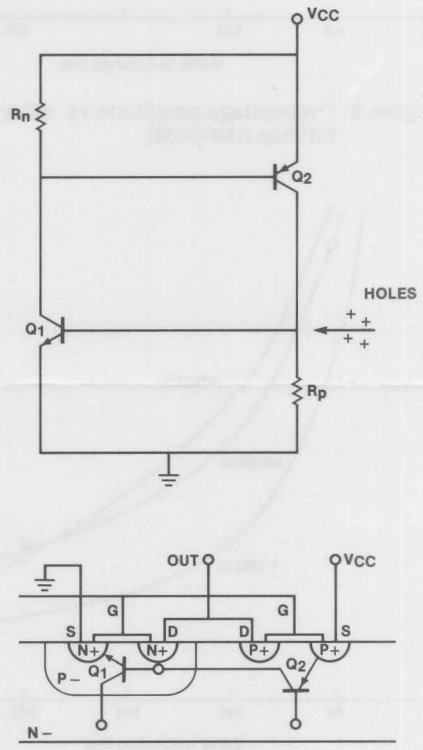


Figure 4: The two transistor SCR model

series bulk silicon RAMs, when now manufactured with the newer RAMs, have exhibited tendencies to latchup in cases where input over-voltage is not monitored closely. For this reason, until special engineering and processing techniques can be developed to eliminate the latchup problem completely, greater care than ever must be taken to avoid triggering conditions.

Some Solutions

Latchup solutions can be divided into two types—those intended to eliminate over-voltages at inputs due to transient or overdriving effects, and those intended to eliminate over-voltages due to power supply switching procedures. The former are "fixes" to system hardware; the latter simply recommendations for software-controlled timing.

[2]Memoire, RAM CMOS. Centre National D'Etudes Spatiales (France). Data compiled May 1980.

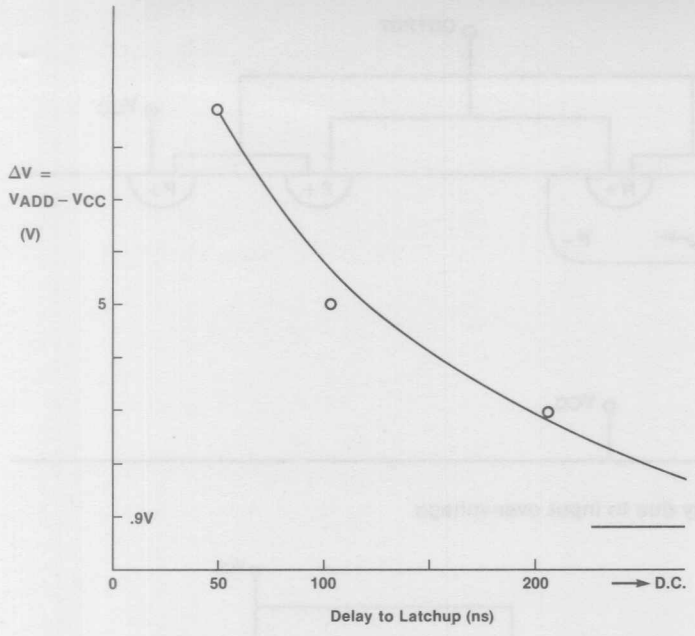


Figure 5: Over-voltage amplitude vs. delay to latchup (IM65X08)

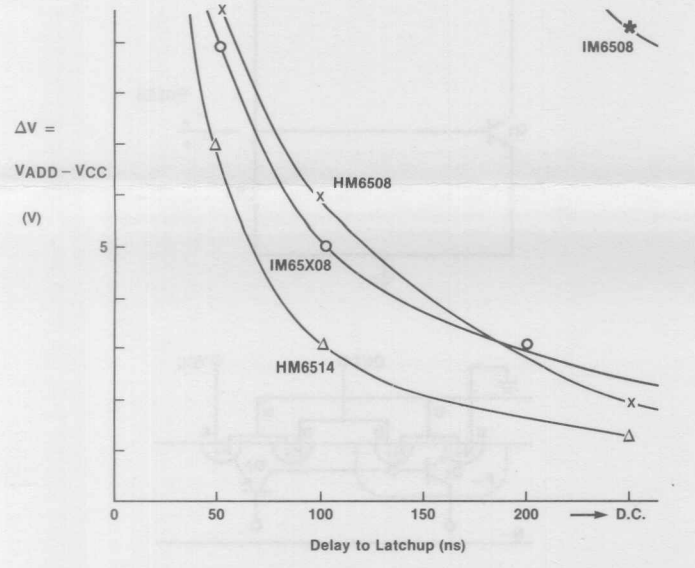


Figure 6: Over-voltage amplitude vs. delay to latchup (4 RAMs)

Transients and Overdriving

Transient over-voltages at chip inputs are particularly difficult to detect and correct. Often they are not repeatable, and without the potential for repetition they can be hard to study. In many cases transients cannot be identified with particular latchup events, but can only be assumed to be triggering influences. If latchup is experienced in a given system and cannot be accounted for otherwise, this assumption should be made, and transients corrected for accordingly.

Fig. 7 shows three variations of a basic technique used to prevent forward bias of the internal input protection diode. In 7a, a common pull-up resistor is diagramed; a good idea for more than one reason, as will be seen in the next section. This configuration has the advantage of requiring an IR drop across the resistor equal to the forward bias voltage of the diode, before the diode will turn on. Normally voltage transients are not able to drive this current, and latchup triggering is avoided. A drawback of the pull-up resistor is delay introduced to the input signal on downward transitions by the resistor bus-capacitance time constant.

Fig. 7b and c show the use of a Schottky diode and a Germanium diode, respectively. These diodes have lower forward bias voltages than the internal protection diode, and when turned on act as shunts for possible latchup triggering current. Typical bias values are 0.5V for the Schottky diode and 0.2V to 0.3V for the Germanium diode. Diode solutions, while limited by the switching speeds of the diodes used, generally introduce less delay, and can handle the large transients that would build up voltage across the resistor of 7a.

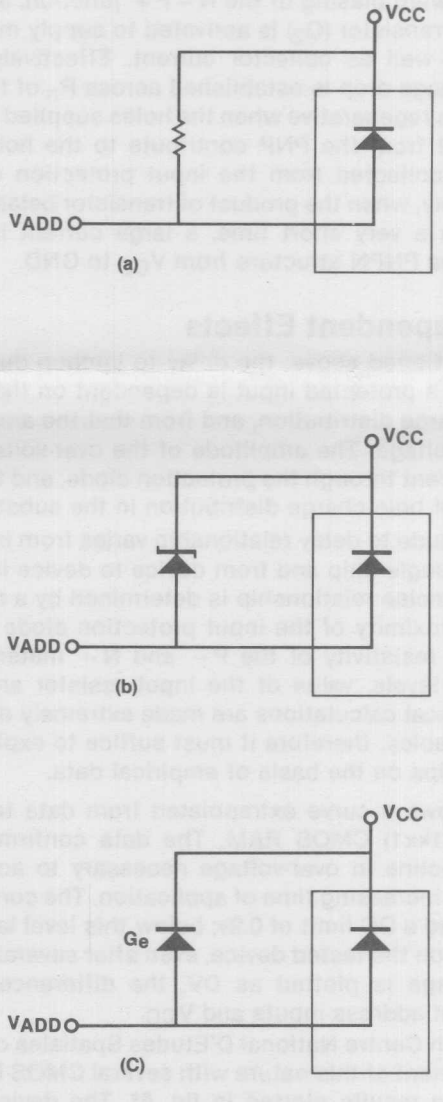


Figure 7 (a): Pull-up resistor;

(b): Schottky diode

(c): Germanium diode

Power Timing and Sequence

Input over-voltages are sometimes accidentally induced during CMOS memory power-up and power-down sequences. When this happens, valid data (i.e. input voltages) can be applied to the memory inputs before the full V_{CC} is reached, or held after V_{CC} is reduced, with latchup resulting. This is possible in cases where the V_{CC} ramps up slowly, or where power is switched from one supply to another, as in battery back-up systems, or multiboard systems with independent power supplies on each board.

Avoidance of latchup during power-up and power-down from OFF conditions is relatively straightforward—simply ensure that the first voltage applied to and the last voltage removed from the memory chip is V_{CC}.

Problems involving transfer of memory V_{CC} from the system supply to a battery supply can be more difficult. This transferring is usually done through a trickle charge/battery isolation circuit (several are common), where interrupted system power turns off a diode or a saturated transistor and leaves the CMOS V_{CC} to be supported by a battery of lower voltage. If the inputs remain at system V_{CC} levels, over-voltage and latchup may inadvertently result. Care must be taken in this case to see that the inputs, if at logic high levels, are decreased to the minimum V_{CC}. This is perhaps best done in anticipation of system power loss (i.e. some sort of software output disable). It is inadequate merely to generate a warning or disabling signal when the battery charge/isolation circuit is activated; by the time it is received and processed, latchup may already have ocurred.

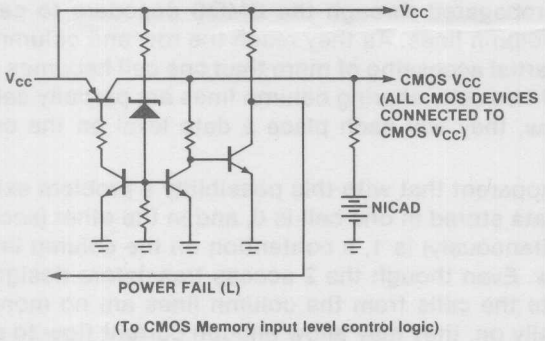


Figure 8: Typical System/CMOS V_{CC} isolation circuit. After system V_{CC} is disabled and battery power asserted, POWER FAIL is activated. However, POWER FAIL may be too late to lower CMOS memory inputs before latchup.

Multiple Address Access and Loss of RAM Data

The Problem

Unique to synchronous CMOS static RAMs is a phenomenon that for want of a better description may be termed "multiple address access". This phenomenon causes loss of stored data at random bit locations, and can be a source of serious concern for system engineers. Though conditions that lead to multiple address accessing will not result in damage to the RAM itself, they may falsely indicate RAM problems. More than a few users have tested and retested RAMs thought to be losing their memory contents, only to find no trace in standard device screening tests of the failures they had experienced during system operation. Many of these data losses were attributable to multiple address access.

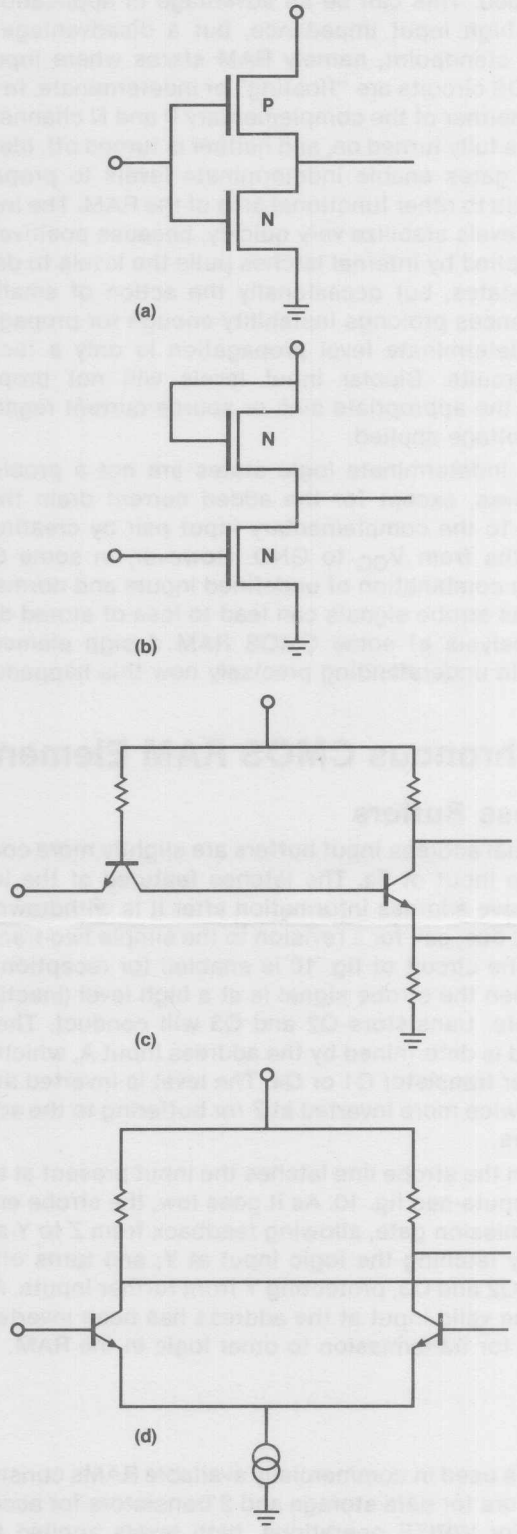


Figure 9: Logic Family Input Circuits

(a) CMOS; (b) NMOS; (c) TTL; (d) ECL

CMOS Inputs

CMOS logic circuits differ from those of other technologies in many ways, one of which is input structure. Like all other MOS circuits they require virtually no input current, as opposed to bipolar inputs which will operate only if current sink (TTL) or source (ECL) capacity

is provided. This can be an advantage in applications requiring high input impedance, but a disadvantage from another standpoint, namely RAM states where inputs to the CMOS circuits are "floating" or indeterminate. In these states, neither of the complementary P and N channel transistors is fully turned on, and neither is turned off. Identical internal gates enable indeterminate levels to propagate, from node to other functional area of the RAM. The indeterminate levels stabilize very quickly, because positive feedback applied by internal latches pulls the levels to defined 0 or 1 states, but occasionally the action of small chip capacitances prolongs instability enough for propagation. This indeterminate level propagation is only a factor in MOS circuits. Bipolar input levels will not propagate without the appropriate sink or source current regardless of the voltage applied.

Internal indeterminate logic states are not a problem of themselves, except for the added current drain they introduce to the complementary input pair by creating current paths from V_{CC} to GND. However, in some CMOS RAMs, a combination of undefined inputs and normal synchronous strobe signals can lead to loss of stored data. A brief analysis of some CMOS RAM design elements is helpful in understanding precisely how this happens.

Synchronous CMOS RAM Elements

Address Buffers

The actual address input buffers are slightly more complex than the input of 9a. The latches featured at the inputs, which save address information after it is withdrawn from the data bus, call for a revision to the simple two-transistor input. The circuit of fig. 10 is enabled for reception of inputs when the strobe signal is at a high level (inactive). In this state, transistors Q2 and Q3 will conduct. The level received is determined by the address input A, which turns on either transistor Q1 or Q4. The level is inverted at node Y, and twice more inverted at Z for buffering to the address decoders.

A low on the strobe line latches the input present at the address inputs-see fig. 10. As it goes low, the strobe enables a transmission gate, allowing feedback from Z to Y and effectively latching the logic input at Y; and turns off transistors Q2 and Q3, protecting Y from further inputs. At this point the valid input at the address has been inverted and latched for transmission to other logic in the RAM.

Cells

The cells used in commercially available RAMs consist of 4 transistors for data storage and 2 transistors for access to them. For WRITE operations, high levels applied to the gates of N – channel transistors Q9 and Q10 in fig. 11 select a cell by row, and the source of either Q9 or Q10 is pulled low, depending on whether data to be written is a 1 or a 0, to select a cell by column. Access to the cell by active voltage on the column lines sets the 4-transistor cell to one of two latch states. When the active voltage is removed, the cell is again isolated and the state maintained by the cross-coupled feedback arrangement. For READ operations, row select is the same, but voltages at the sources of Q9 and Q10 are measured by a sense amplifier instead of being forced to the desired level.

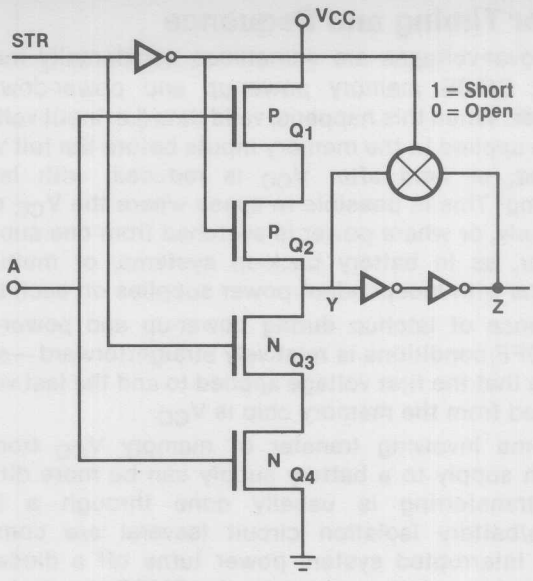


Figure 10: CMOS RAM Address Buffer

Multiple Address Access

Floating or disconnected address inputs, appearing at A of fig. 10, fail to turn input transistors Q1 and Q4 fully on or off. The direct result is the introduction of indeterminate logic levels to the internal CMOS circuits. Node Y becomes indeterminate, and propagates to Node Z and beyond before it is stabilized by positive feedback.

With the normal activation of the strobe, these false levels are propagated through the CMOS decoders to cell row and column lines. As they reach the row and column lines the partial accessing of more than one cell becomes possible. If two cells sharing column lines are partially selected by row, they will each place a data level on the column lines.

It is apparent that with this possibility a problem exists. If the data stored in one cell is 0, and in the other (accessed simultaneously) is 1, a contention on the column line will follow. Even though the 2 access transistors designed to isolate the cells from the column lines are no more than partially on, they may allow enough current flow to permit the writing of a 1 to a cell containing 0, or the reverse. In this way, accidental multiple address access may cause transformation and therefore loss of stored data.

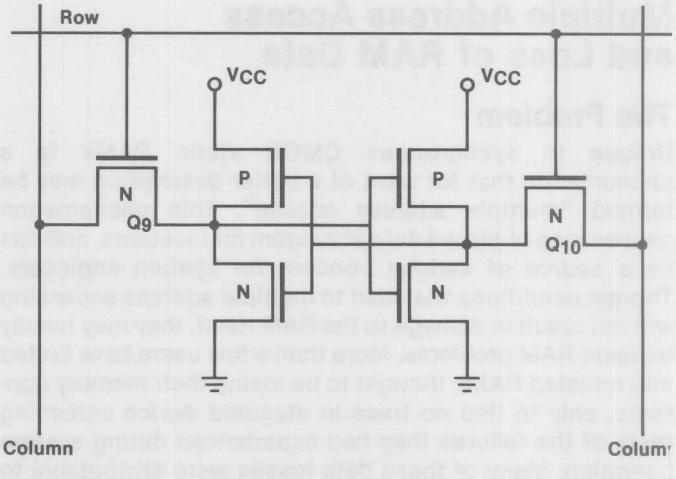


Figure 11: CMOS RAM Cell

Some Solutions

Multiple address access most frequently occurs during RAM power-down periods. During these periods, inputs to the RAMs are typically left at high impedance levels; the outputs that drive them having been disabled or disconnected in some manner. At the same time, the RAMs are continually strobed or accessed, a procedure which does not at first appear harmful, assuming the RAM outputs themselves are disabled and prevented from driving the data bus.

To prevent multiple address access, one of these conditions must be eliminated. Either the address inputs must be held at defined logic levels, within the device specification limits, or the RAM must not be strobed during this time. Though use of both procedures may be best, if only one is to be used the former is more desirable. If the address lines are well-defined current will not flow through the input buffers, and system power, often at a premium in power-down periods, will be conserved.

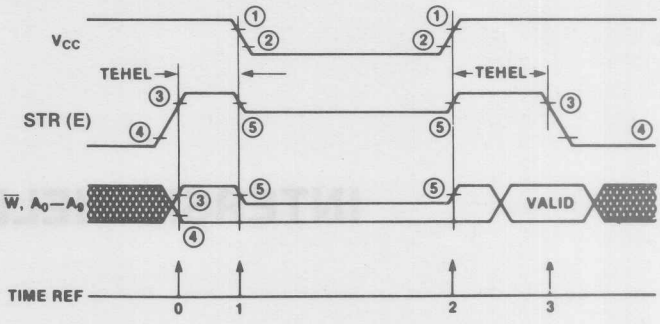
The address lines can be made determinate with pull-up resistors, as diagrammed in fig. 7a; resistor values of about 10k are usually sufficient. Precise values can be chosen according to the power/speed tradeoff; higher resistance decreases static current flow, but slows normal input response on downward transitions, in conjunction with the bus capacitance.

Should pull-up resistors be an unacceptable alternative, there remains the choice of disabling the strobe, at a high level until the next RAM operation. If the strobe signal is provided from a controlling device, such as a microprocessor with a synchronous data bus, it may continue to change levels despite the inactive RAM state. In this case the RAM strobe input must be decoded with chip select signals normally used to control the output buffers, via external logic.

Recommended Power-Down Timing

A typical RAM power-down sequence is diagrammed in fig. 11. The strobe or E signal in this sequence is gated to a high level at T = 0 and is held there for a specified minimum time period before voltage levels decrease to those of the inactive state (T = 1). Address and other input lines are also pulled high (or low) at this time. The same sequence is observed in reverse, when the RAM is returned to system power and operating conditions.

Of particular note in this timing sequence is the provision for a minimum strobe inactive time, (designated TEHEL in fig. 12). If this timing requirement is not met, internal timing functions are interrupted, and data in RAM cells can be lost in much the same manner as multiple address access. For example, it is apparent that the strobe line must be inactive (high) long enough to disable the transmission gate and allow new data to be latched in the address buffer (see fig. 10). If TEHEL is violated, internal data contention, indeterminate levels and loss of RAM data may result.



① 4.5V
② V_{CCDR} (5.5V ≥ V_{CCDR} ≥ 2.0V)
③ V_{IH} (V_{CC} + 0.3V ≥ V_{IH} ≥ V_{CC} − 2.0V)
④ V_{IL} (0.8 ≥ V_{IL} ≥ GND—0.3V)
⑤ V_{IHDR} (V_{CC} + 0.3V ≥ V_{IHDR} ≥ V_{CC} − 2.0V ≥ 2.0V)

Figure 12: CMOS RAM Power-down timing

REFERENCES

1. SZE, S.M., *Physics of Semiconductor Devices,* Wiley and Sons, 1969
2. Ochoa, A., Dawes, W., and Estreich, D., "Latch-up Control in CMOS Integrated Circuits", *IEEE Transactions on Nuclear Science*, Vol. NS-26, No. 6, p. 5065.
3. Flood, J. and Pujol, H.L., "Guide to Better Handling and Operation of CMOS Integrated Circuits", RCA *COS/MOS Integrated Circuits,* 1977
4. Centre National D'Etudes Spatiales, France. Data from study on CMOS Integrated Circuits, compiled May 1980

INTERSIL FIELD SALES OFFICES

CALIFORNIA

1272 Forgewood Avenue
Sunnyvale, California 94086
Tel: (408) 744-0618
TWX: 910-339-9260

400 Oceangate Suite #1102
Long Beach, CA 90802
Tel: (213) 436-9261
TWX: 910-341-6829

COLORADO

5 Parker Place, Suite #351
2600 S. Parker Road
Aurora, Colorado 80014
Tel: (303) 750-7004
TWX: 910-320-2982

FLORIDA

Hollywood 95 Office Park
2700 N. 29th Avenue
Building #2, Suite #204
Hollywood, Florida 33020
Tel: (305) 920-2442
TWX: 510-954-9819

ILLINOIS

201 Ogden Avenue, Suite #230
Hinsdale, Illinois 60521
Tel: (312) 986-5303
TWX: 910-651-0859

MASSACHUSETTS

2 Militia Drive, Suite 12
Lexington, Massachusetts 02173
Tel: (617) 861-6220
TWX: 710-326-0887

MINNESOTA

6550 York Avenue, South, Suite #307
Minneapolis, Minnesota 55435
Tel: (612) 925-1844
TWX: 910-576-2780

NEW JERSEY

560 Sylvan Avenue
Englewood Cliffs, New Jersey 07632
Tel: (201) 567-5585
TWX: 710-991-9730

OHIO

228 Byers Road
Miamisburg, Ohio 45342
Tel: (513) 866-7328
TWX: 810-473-2981

TEXAS

10300 N. Central Expwy.
Suite 225-III
Dallas, TX 75231
Tel: (214) 369-6916
TWX: 910-860-5482

CANADA

338 Queen Street East, Suite #208
Brampton, Ontario L6V 1C4
Tel: (416) 457-1014
TWX: 610-492-2691

INTERSIL

CORPORATE HEADQUARTERS
10710 N. Tantau Avenue
Cupertino, CA 95014
Tel: (408) 996-5000
TWX: 910-338-0171

NORTHERN AND EUROPEAN HEADQUARTERS
9th Floor
Snamprogetti House
Basing View
Basingstoke RG21 2YS
Hants, England
Tel: (0256) 57361
TLX: 858041 INTRSL G

SOUTHERN EUROPEAN HEADQUARTERS
Bureau de Liaison
217, Bureaux de la Colline,
BAT. D (2e Etage)
92213 Saint-Cloud Cedex
France
Tel: (1) 602.58.98
TLX: DATELEM 204280F

CENTRAL EUROPEAN HEADQUARTERS
Bavariaring 8
Concordiahaus
8000 München 2
West Germany
Tel: 89/539271
TLX: 5215736 INSLD

FAR EAST HEADQUARTERS (EXCEPT JAPAN)
c/o S.S.I. Far East Ltd.
Suite 201, Austin Centre
21 Austin Avenue
Tsim Sha Tsui Kowloon
Hong Kong
Tel: 3-672112-3
TLX: 86496 SSI HX

4-81-00A LASK NAHT